## Identifying disease-centric subdomains in very large medical ontologies

Milian, K.; Aleksovski, Z.; Vdovjak, R.; ten Teije, A.C.M.; van Harmelen, F.A.H.

***published in***
Knowledge Representation for Health-Care: Data, Processes and Guidelines, AIME 2009, Workshop KR4HC 2009, Revised Selected and Invited Papers

2010

***DOI (link to publisher)***
10.1007/978-3-642-11808-1_5

***document version***
Early version, also known as pre-print

# Identifying Disease-Centric Subdomains in Very Large Medical Ontologies: A Case-Study on Breast Cancer Concepts in SNOMED CT. Or: Finding 2500 Out of 300.000

Krystyna Milian[1], Zharko Aleksovski[2], Richard Vdovjak[2], Annette ten Teije[1], and Frank van Harmelen[1]

[1] Vrije Universiteit Amsterdam
krystyna.milian@few.vu.nl
[2] Philips Research
zharko.aleksovski@philips.com

**Abstract.** Modern medical vocabularies can contain up to hundreds of thousands of concepts. In any particular use-case only a small fraction of these will be needed. In this paper we first define two notions of a disease-centric subdomain of a large ontology. We then explore two methods for identifying disease-centric subdomains of such large medical vocabularies. The first method is based on lexically querying the ontology with an iteratively extended set of seed queries. The second method is based on manual mapping between concepts from a medical guideline document and ontology concepts. Both methods include concept-expansion over subsumption and equality relations. We use both methods to determine a breast-cancer-centric subdomain of the SNOMED CT ontology. Our experiments show that the two methods produce a considerable overlap, but they also yield a large degree of complementarity, with interesting differences between the sets of concepts that they return. Analysis of the results reveals strengths and weaknesses of the different methods.

**Keywords:** identifying ontology subdomain, disease related concepts, ontology subsetting, mapping medical terminologies, seed queries, medical guidelines.

## 1 Introduction

Large medical ontologies such as SNOMED CT [1] contain hundreds of thousands of clinical concepts usually organized in a hierarchy and interconnected by domain specific relations, together representing the explicit semantic knowledge describing a medical field. Such knowledge can be of great help when developing intelligent clinical decision support systems that focus on reasoning about patient data within a certain disease domain. A disease-specific, richly annotated

[1] http://www.ihtsdo.org/snomed-ct/

D. Riaño et al. (Eds.): KR4HC 2009, LNAI 5943, pp. 50–63, 2010.

semantic subdomain is also an important element in the process of overcoming the frequent problem of lexical heterogeneity between the concepts occurring in the patient data and those from the applicable clinical guidelines. However, identifying a *disease-centric subdomain* of a large medical ontology is not a trivial task. The relevant concepts are seldom to be found under one sub-branch of the ontology, instead they are usually scattered in various branches representing different facets of the domain coverage, e.g. clinical findings, procedures, anatomic regions, etc.

In this paper we describe a study on the identification of SNOMED CT concepts related to breast cancer. We compare results of two different methods: *(i)* The *seed query method* from [1] was used for extraction of concepts that are unique to breast cancer. *(ii)* The so-called *guideline-based method*, consisting of a manual mapping between SNOMED CT concepts and the important concepts from the Dutch national breast cancer guidelines, was used for the identification of those concepts that are relevant with respect to breast cancer.

Our experiments show that the two methods produce a considerable overlap, but they also yield a large degree of complementarity, with interesting differences between the sets of concepts that they return. The size of the identified subdomains is considerably smaller than that of the whole medical ontology (between 0.1%-1%), making the reasoning as well as the maintenance task of such subdomain much more feasible.

The paper is structured as follows: Section 2 introduces different notions of relevancy in subdomains of a medical ontology, and puts forward the main hypothesis of the paper. Section 3 and 4 introduce our two different subdomain-selection methods: the seed query method in section 3 and the guideline-based method in section 4. Section 5 compares and analyses the results. Section 6 presents related work. Section 7 summarizes the findings and presents the concluding remarks.

## 2 Two Types of Disease-Centric Subdomains

Before investigating methods for identifying disease-centric subdomains from a large ontology, we must first define what we mean by such a subdomain. For the purpose of this paper we will set our own definitions. Presented below methods are not based on any *a priori* modularization of the ontology, but they identify subdomains that are specific for any particular use of a vocabulary.

*Definitions:* We distinguish two kinds of disease-centric subdomains, namely *relevant subdomains* and *key subdomains*, which consist of relevant concepts and key concepts respectively. The notions of "relevant concepts" and "key concepts" are each defined as follows:

**Relevant concept**: A concept C is a *relevant concept* for a disease D if clinical guidelines for D state that it influences a decision on the diagnosis or treatment of D.

An example of a concept that is relevant to breast cancer is "pregnancy": datasources about breast cancer (such as guidelines, patient-records, textbooks, etc) often contain the concept "pregnancy" because certain treatments (e.g. chemotherapies) are ruled out for pregnant women.

However, the converse is not the case: not any document containing the concept "pregnancy" is likely to be about breast cancer. To capture this, we define a second notion:

**Key concept**: A concept C is a *key concept*[2] for a disease D if the occurrence of C in a datasource S means that S is conclusively about D.

An example is the concept "malignant neoplasm of breast". Any key concept is of course a relevant concept, but not vice versa.

*Hypothesis:* Our hypothesis is that the seed query method (described in section 3), when seeded properly, will identify only key concepts, while the manual guideline-based method (described in section 4) will identify relevant concepts. From the above definitions, this hypothesis also implies that the seed query results should be contained in the guideline-based results.

*Choice of dataset:* In this paper, we focus on breast cancer as our clinical domain both because of its prevalence and the highly progressed state-of-the-art in diagnoses and treatment, which is expected to involve a relatively rich vocabulary and thus presents an interesting use-case. We concentrate on SNOMED CT as our main ontology, mainly because of its high adoption and a broad clinical coverage, containing more than 300.000 concepts. Besides applying both methods to the breast cancer domain in SNOMED CT, we also apply the seed query method to three other very large ontologies, namely to NCI, MeSH and ICD10. We do this to verify the consistency of our results. The precise use of these ontologies is described in next section. Also applying the manual guideline-method to these ontologies would have been prohibitively expensive.

## 3 Seed Query Method to Find Key Concepts

*Method.* The seed query method, originally published in [1], is a combination of a lexical and a structural approach.

It takes a list of combinations of some of key concepts (the so-called "seed queries"), which serve as prior knowledge, to find an initial set of domain specific, in this case breast cancer related concepts through lexical mapping to the concepts in the ontology. This set is then expanded through the hierarchical structure of the ontology, and through the structure of UMLS (Unified Medical Language System[3]) metathesaurus. Given a set of seed queries, the process is completely automatic, ensuring repeatability of the extraction. It also allows for gradual improvement by adjusting the initial set of seed queries.

[2] "key" is inspired by the database notion of the same name

[3] http://www.nlm.nih.gov/research/umls/

In more detail, the seed query method proceeds in three steps: *(i) Query matching* which uses the concept's names, *(ii) Subconcept expansion* based on the hierarchical structure of the ontologies, and *(iii)* UMLS expansion which uses the UMLS metathesaurus. The three steps in this method are sequential, increasing the set incrementally, each step produces set of concepts which is passed as input to the next step. The third step produces the final result of the method. Next, we elaborate each of the steps, and also present it as a pseudo-code algorithm.

*Query matching* uses a list of seed queries to find concepts from the subdomain by trying to lexically match the queries to each concept from the ontology. The lexical match was not sensitive to letter capitalization, and in addition, Porter's stemmer algorithm [11] was used to normalize the words before comparison. Such queries consist of keywords or combinations of keywords which are specific to the subdomain, and when a concept lexically matches to some of these queries, it can be considered part of the subdomain. The algorithm for query matching is shown in Figure 1. It is applied on each of the four candidate ontologies separately.

*Subconcept expansion* expands the set of concepts produced in the first step by including their subconcepts. Each ontology generally organizes the concepts in a hierarchy through IS-A relations among them (e.g. Breast cancer IS-A Cancer). These relations were used to find all the subconcepts of the concepts found in the first step. This process was done exhaustively, transitively adding the subconcepts of the newly found concepts as well, until no new concepts could be added. The algorithm for subconcept expansion is shown in Figure 2. It is separately applied on each set obtained in the first step.

UMLS *expansion* uses UMLS to further increase the set produced in the second step. UMLS assigns a unique identifier to every concept from every ontology integrated in it, and if two concepts have the same identifier then they mean the same thing. Suppose two arbitrary concepts $A_1$ and $A_2$ from two ontologies $\mathsf{ONT}_1$ and $\mathsf{ONT}_2$ respectively, are assigned the same identifier in UMLS. Now, if $A_1$ is found as key concept in the first two steps for the ontology $\mathsf{ONT}_1$ and $A_2$ is *not* found as key concept for the ontology $\mathsf{ONT}_2$ in the first two steps, then $A_2$ can be added as a key concept for the ontology $\mathsf{ONT}_2$, thus expanding the set of key concepts for the ontology $\mathsf{ONT}_2$. This way of expanding the sets of key concepts is the third step of the method. It is done exhaustively, for every concept and every pair of ontologies used in the experiment. The algorithm for UMLS expansion is given in Figure 3. It is applied on the four sets of key concepts obtained in the second step.

*Results.* The breast cancer-centric subdomain of SNOMED CT (containing only key concepts for breast cancer) was extracted using the method described above.

We seeded the method with a hand-crafted list of breast cancer seed queries, shown in Table 1. After starting with a small number of key concepts, and iteratively adding seeds, we observed that after a small number of concepts the results stabilise, and no longer grow when adding further key concepts as seeds. This process has up to now been informal, and would merit a more detailed study in its own right.

*The resulting set of matched concepts is empty in the beginning*
1 subdomain $:= \emptyset$
*Lexically matching the concepts from the ontology to the query list*
2 **for** each query $Q \in$ list of queries **do**
3 **for** each concept $C \in \mathcal{C}^{\mathsf{ONT}}$ **do**
4 **if** LexicalMatch$(C, Q)$ **and** $C \notin$ subdomain **then**
5 subdomain $:=$ subdomain $\cup \{C\}$

**Fig. 1.** Step one: Query matching

*Add all the subconcepts to the concepts in* subdomain
1 **while** adding new concepts in subdomain is possible **repeat**
2 **for** each concept $X \in$ subdomain **do**
3 **for** each concept $Y \in \mathcal{C}^{\mathsf{ONT}}$ **do**
4 **if** $Y \subseteq X$ **and** $Y \notin$ subdomain **then**
5 subdomain $:=$ subdomain $\cup \{Y\}$

**Fig. 2.** Step two: subconcept-based expansion

*Expanding each of 4 result sets through* UMLS
1 **for** any $\mathsf{ONT}_p, \mathsf{ONT}_q \in \{\mathsf{SNOMED\ CT}, \mathsf{NCI}, \mathsf{MeSH}, \mathsf{ICD10}\}$ , where $p \neq q$ **do**
2 **for** each concept $X \in \mathsf{subdomain}_p$ **do**
3 **for** each concept $Y \in \mathsf{subdomain}_q$ **do**
4 **if** UMLS : $X \equiv Y$ **and** $Y \notin \mathsf{subdomain}_p$ **then**
5 $\mathsf{subdomain}_p := \mathsf{subdomain}_p \cup \{Y\}$

**Fig. 3.** Step three: UMLS-based expansion

Besides SNOMED CT, the method was applied on three other ontologies: NCI[4] - a vocabulary for annotating medical documents primarily cancer related, MeSH[5] - a vocabulary for scientific literature annotation and ICD10[6] - a classification of diseases. The ontologies were used as extracted from the UMLS 2008AA version.

The results of applying the seed query method are shown in Table 2. The table shows that only a fraction of the entire ontology (much less than 1%) are key concepts for a disease such as breast cancer. It also shows that most of the results are actually found in the first phase. This is reasonable: most of the concepts are very specialized and are hence leafs in the ontologies. Finally, it is interesting to see that the most specialised ontology (the oncology-specific NCI) has the highest hit-rate of key concepts, and the most general and wide ranging ontologies (MeSH and SNOMED CT) have the lowest hit-rates.

[4] http://nciterms.nci.nih.gov
[5] http://www.nlm.nih.gov/mesh
[6] http://www.ahima.org/icd10

**Table 1.** Seed queries used to extract the breast cancer subdomain

1. Breast cancer
2. Breast carcinoma
3. Microcalcification
4. Mammary carcinoma
5. Lobular carcinoma
6. Ductal carcinoma
7. Mastectomy
8. Paget breast
9. HER2/neu
10. HER-2
11. BRCA

**Table 2.** Results of applying the seed query method on the four ontologies: incremental results are reported after each step (full method = after step 3)

| Ontology | size of ontology | number of concepts extracted after step 1 | after step 2 | after step 3 | % of full ontology |
|---|---|---|---|---|---|
| SNOMED | 308,677 | 198 | 271 | 279 | 0.09% |
| NCI | 62,969 | 358 | 388 | 399 | 0.63% |
| MeSH | 282,425 | 105 | 120 | 129 | 0.05% |
| ICD10 | 11,529 | 5 | 5 | 12 | 0.10% |

## 4 Mapping of Guidelines to Find Relevant Concepts

In this method, we used clinical guidelines a source of information about domain related concepts, in order to identify a disease-centric subdomain of an ontology. Medical guidelines describe recommendations and conclusions regarding proper treatment based on scientific evidence. They aim to reduce the growing gap between knowledge and the actual practice. In our research, we used breast cancer guideline developed by the joint initiative of the Dutch Institute for Health care Improvement (CBO) [3].

From formalised models of the guideline [8] we extracted the names of all treatment plans, as well as all parameters describing patient data and their possible values in case of enumerated types. The parameters either specify plan preconditions and intentions or data that can be requested from external sources during guidelines execution. We mapped extracted concepts manually and had it verified by medical expert. Then we used the obtained mapping as a gold standard to compare with the results which could be produced by automatic mapping tool, the MetaMap [2].

*Practical experiences with manual mapping.* The main challenges of mapping concepts extracted from the guidelines to SNOMED CT concepts were searching among the hundreds of thousands of SNOMED CT concepts for the equivalences. Mapping required understanding the meaning of concepts used in the

guidelines and knowing the exact context where they were used. After the initial mappings were identified, we consulted with our clinical expert and made adjustments where necessary. Below we illustrate some of the difficulties which we encountered.

In many cases guidelines and SNOMED CT use different terminology to express the same information. 'Axillary-node-dissection-proper' used in the guidelines and 'Excision of axillary lymph node' defined in SNOMED CT are an example of such case. Finding corresponding concepts was done using key words or using synonyms found in medical dictionaries. In cases where both approaches failed, we checked the context in the guidelines or looked for an explanation of concepts in other resources. This applied in the case of abbreviations as well as full phrases.

On the other hand finding an exact lexical match can be sometimes misleading. Such a situation was encountered when the plan 'Mastectomy' was analyzed. In the guidelines it covers the plan 'Mastectomy-proper' and also other procedures such as 'Radiotherapy-chest-wall' and 'Breast-reconstruction'. Hence the plan 'Mastectomy-proper' rather than 'Mastectomy' should be mapped to the SNOMED CT concept 'Mastectomy'. Therefore knowing the context was necessary.

Differences in granularity and abstraction level caused most of the missing matches. This issue appears mostly in the case of multiterms concepts, which are commonly used in the guidelines. Examples of such compound concepts are therapy + drug e.g. anthracycline-chemotherapy-manual, or therapy + drug + number of repetition e.g. six-courses-anthracycline-chemotherapy. Multiterms concepts are also used to define the intentions of therapies, for example 'elimination-distant-metastases'. Such specific concepts turned out to be very unlikely to be found in SNOMED CT ontology.

In a few cases even the large SNOMED CT ontology is not expanded enough yet. For example, SNOMED CT contains no concept corresponding to the parameter 'patient-prefers-bct', describing the patients preference of breast conserving treatment over mastectomy.

All these points above show that the method of obtaining relevant subdomains by mapping from guidelines is essentially a manual operation that cannot easily be automated. Results of manual mapping are significantly better, our early work in this domain ([12]) also corroborate this.

*Results of the manual mapping.* We found around 60 exact matches (matches with the same meaning but not necessarily the same name) out of all 150 parameters extracted from the guidelines. In the case of treatment procedures, we found around 40 exact matches out 190 procedures, and 40 matches, where SNOMED CT concepts have a close but more general meaning. The missing matches are caused by the reasons mentioned above.

*Benchmark against MetaMap.* In order to verify that manual mapping is indeed necessary we tested the applicability of MetaMap tool for the purpose of our research. MetaMap is a program developed at the National Library of Medicine

to map biomedical text to the Metathesaurus [2]. It combines computational linguistic techniques with symbolic, natural language processing. MetaMap performs mapping in five main steps:

1. Parsing. The entire text is parsed, and divided into simple phrases using the SPECIALIST minimal commitment parser [9] which produces a shallow syntactic analysis of the text.
2. Variant Generation. In second step for each phrase are generated variants, using SPECIALIS lexicon and a database of synonyms, including all acronyms, synonyms, derivational and spelling variants of the given phrase.
3. Candidate Retrieval. Further the algorithm retrieves the set of all Metathesaurus strings, containing at least one of the variants.
4. Candidate Evaluation. Retrieved candidates from Metathesaurus are used to generate the mappings, which are evaluated using a linguistically principled evaluation function consisting of a weighted average of metrics measuring centrality, variation, coverage and cohesiveness. Then the list is ordered according to calculated scores.
5. Mapping Construction. Final mapping are constructed by combining candidates involved in disjoint parts of the phrases, and evaluated using the same scoring function. The mapping with the highest score is the best proposal of MetaMap.

We tested MetaMap on the same set of parameters and treatment plans extracted from the Breast Cancer guidelines. We compared the results with the results obtained by the manual mapping experiment. For each concept extracted from the guidelines, we checked whether its corresponding SNOMED CT concept, identified during manual mapping, is in the list of candidates proposed by MetaMap. In order to avoid ambiguity, and include equivalent mapping of different synonyms, we used for the comparison UMLS identifiers instead of concept names. It was possible due to the fact that SNOMED CT is included in the UMLS Metathesaurus. We tried different settings options to gain the deeper insight of MetaMap possibilities, including 'Term processing' and 'Ignore stop phrases'. In 'Term processing' mode input text is not divide into simple phrases but considered as a whole, which seems to be more adequate in the case of mapping concepts, which are most commonly multiword concepts.

The biggest overlap between the results produced by these two different mapping methods contains 30 out of 190 treatment plans and 16 out of 150 parameters. It was obtained using 'Term processing' mode. When for the comparison were used only the best candidates of MetaMap algorithm, then the numbers of exactly the same mappings decreased to 22 and 14 in case of plans and parameters respectively. Obtained results are summarized in table 3.

The major reason for differences in obtained mapping result are different strategies used for dealing with multiterms concepts. MetaMap proposes list of individually mapped Metathesaurus concepts, whereas we were aiming for finding a single corresponding concept with the closest meaning. For example MetaMap suggests for the concept 'Tumour negative excision margins' following mapping : 'Tumor excision NOS', 'Negative', 'Margin (Marginal)'. Manual

**Table 3.** Comparison of results of identified SNOMED CT concepts obtained using different mapping strategies

| Mapping strategy | Identified parameters | Identified plans |
|---|---|---|
| MetaMap (all) | 16 (10%) | 30 (15%) |
| MetaMap (first) | 14 (9%) | 22 (12%) |
| Manual | 60 (40%) | 80 (41%) |

browsing of the ontology and awareness of the application context let us identify the actual corresponding SNOMED CT concept 'Breast surgical margin not involved by tumour'. Automatic identification of such lexically unrelated concepts could be possible if for example SNOMED CT contained rich enough list of synonyms.

Results obtained using MetaMap, 15% of plans, and 10% of parameters correctly mapped to single corresponding concepts, clearly show that using mapping tools, which focus on lexical matching, is not sufficient in case of text composed using non-standard terminology, as that provided by SNOMED CT. It confirms our concern that manual mapping is necessary manual exercise in such case.

The set of identified SNOMED CT concepts, obtained by mapping guidelines concepts will be further expanded as described below.

*Results of the expansion steps.* In section 3, seed queries were used for the lexically querying for matching concepts. In the guideline-based method, this step is performed more semantically, namely by manually mapping the parameters and procedures of the guideline. In both cases, this first step is followed by subconcept-based expansion (transitively including all subsuming concepts, fig. 2) and UMLS expansion (using UMLS to include equivalent concepts, fig. 3).

Applying these two expansion steps to the results of the first manual mapping step resulted in an expansion from 140 to 2250 concepts. The two expansion steps have a much bigger impact after the manual mapping (from 140 to 2250) than they have after the first step in the seed query method (from 198 to 279). This difference can be explained by the fact that the first step in the seed query method returns mostly very specific SNOMED CT concepts that have very few subconcepts, while the manual mapping also yielded concepts higher in the SNOMED CT hierarchy.

However, also in the manual mapping case, the breast cancer-centric subdomain is again a very small fraction of the entire ontology, namely 0.73 % of the full ontology (308.677 concepts).

## 5 Evaluation of the Two Methods

Our two methods for identifying breast cancer-centric subdomains provided different results. The manual guideline-method found 2250 concepts, against 279 concepts found by the seed query method. Of these 279 concepts, 155 are

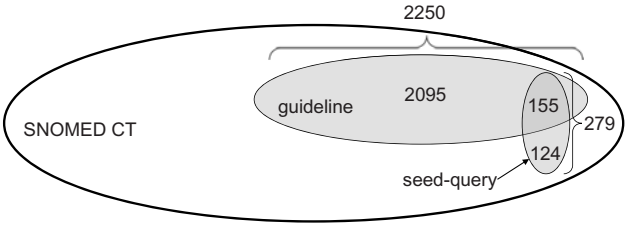


**Fig. 4.** Breast cancer subdomains identified using different approaches

also found by the guideline-method. The inclusion relations are summarised in figure 4.

Unsurprisingly, all 2250 concepts found by the guideline-method are indeed relevant concepts for the breast cancer-centric subdomain, in other words this method has a high precision. This is unsurprising because all concepts are either direct mappings from parameters or procedures in the recommendations of a national breast cancer guideline, or are subconcepts of these concepts.

More interestingly, manual inspection of the 279 seed query results shows that this method has a near perfect precision (i.e. all the concepts it finds are indeed key concepts for the breast cancer-centric subdomain). This confirms the main hypothesis put forward in section 2.

The figure also shows that besides its high precision (finding only key concepts), the seed query method has a rather low recall: it finds less than 10% of the concepts found by guideline-method. This is to be expected since the seed query method is tuned to find only key concepts (instead of finding all relevant concepts). However, inspection of the 2095 concepts that are only found by the guideline-method reveals that there are quite a few key concepts still contained in that set. Hence, even when counting only key concepts, the seed query method has no perfect recall. Examples of obvious concepts that we found missing are "Breast surgical margin involved by tumor", very detailed concepts such "Metastasis in internal mammary lymph nodes with microscopic disease detected by sentinel lymph node dissection but not clinically apparent" and quite a few others.

Finally, and contrary to our prediction, the seed query results are not a subset of the results from the guideline-method. In fact, well over 40% of all seed query results (124) are not found by the guideline-method. Inspecting this set yielded the following explanations for this falsification of our hypothesis:

*Guidelines do not cover diagnostic concepts:* The biggest part of concepts in this group describe breast neoplasm in general, e.g. 'Carcinoma in situ of female breast'. The guidelines are focused on recommendation for treatment of already diagnosed breast cancer, which is malignant. Benign neoplasm is not broadly discussed, since such concepts would be rather covered by diagnosis guidelines.

*Only the guidelines recommendations were used:* Some of those concepts are connected with breast cancer but are not included in the recommendations, the

only part of the guidelines which was formalized. For example recommendation do not mention treatment procedures for male breast cancer, whereas concepts like 'Carcinoma in situ of male breast' or 'Carcinoma in situ of areola of male breast' were identified by seed queries. Moreover guidelines predominantly focus on ductal carcinoma as it is the most prevalent disease. Other types such as lobular carcinoma were only mentioned marginally - and did not occur in the formalized version.

*The guidelines do not mention procedures that vary between hospitals:* In The Netherlands, some hospitals employ special oncology nurses for home care of patients, others don't. The national guidelines do not discuss procedures for which there is an accepted high local variance between hospitals.

Between them, these reasons would remove a substantial part of the outlying 124 concepts, although we are currently not able to determine the exact number.

## 6 Related Work

Identification of a disease centric subdomain out of a large medical ontology to some extent resembles the problem of ontology modularization [4] which is applied in the context of combining existing ontologies by importing relevant modules. While not exactly the same[7], our notion of a subdomain can be compared to the ontology module as defined in [4] and hence we consider the papers presented below as related work.

Existing methods in the literature often rely on an a priori modularization of the vocabularies. These are typically based on some notion of semantic distance, or on the connectivity-graph of the ontology [6,13]. Such methods providing uncustomized modularization do not fulfill the requirement which we are aiming to meet, identifying subdomains specific for a particular use of a vocabulary. However, the methods which create partitioning of an ontology based on a given signature can be an alternative to the presented here seed query method. We will have a closer look to some of them. Generally, modularization techniques are divided into prescriptive and analytic. In prescriptive approach, the user explicitly states what is in or outside of the module, which requires the changes in the syntax and semantics of the language. In [4] one can find many arguments against it. The authors stress the fact that consequently whole infrastructure, OWL reasoners and parsers have to be changed as well. This approach leads to the tight, non-standard solutions, which severely restricts the reusability by other organization. Therefore we will focus on description of techniques based on analytic approach, where ontologies are defined using standard syntax and semantics of OWL, and ontology tools provide modularization services. In [4] they are evaluated according to the aspect of module correctness (any inference deduced from the module should be deduced from the original ontology), module

[7] In our case, we do not impose all formal properties that a module has as it is not necessary in our target application.

completeness (a module should contain all relevant information, so the user can not recognize that not whole ontology is imported) and module minimality ( a module should be as small as possible).

One of analytic, ad hoc solution can be produced using PromptFactor algorithm [10]. It extracts a fragment of an ontology, based on a given signature. Created modules contain axioms that are mentioned in that signature and are further expanded with other concepts mentioned in those axioms until a fixed point is reached. The algorithm has been evaluated in [5], where authors prove that it is not always complete and creates modules larger than those, created by other algorithms that can guarantee completeness.

CEL and MEX are algorithms which work only with tractable fragments of OWL, the EL family of DL. This restriction is not problematic in case of SNOMED CT ontology, but NCI thesaurus and GALEN are beyond expressiveness of EL. The CEL reasoner [14] provides modularization technique based on connected reachability. Reachability can be expressed by a graph, where nodes are labeled with concepts from the ontology and edges are labeled with axioms. The modules contain the concepts themselves and the concepts and axioms from labels of connected nodes. They are guaranteed to be complete. MEX [7] can be applied only for acyclic EL ontologies, it generates minimal modules, smaller then other more generic algorithms.

Locality based algorithms are seen as most promising ones. Informally axiom is local if it does not change the meaning of concepts if included in the module. Changes of meaning are recognized differently according to the chosen locality type, e.g. axiom is top-local for a class if it does not define a new subclass for the concept. Basing on the application one can choose top or bottom locality, to be able to effectively generalize or refine set of identified axioms. Produced modules are proven to be correct and complete and the empirical analysis described in [5] attests the better approximation of minimal modules then other known algorithms.

## 7 Summary and Conclusions

*Summary.* Medical vocabularies are typically very large, containing up to hundreds of thousands of concepts. However, for any particular usage of such vocabularies only a small fraction of the concepts will be needed. In our example use-case, the breast cancer-centric subdomain of SNOMED CT is at most 1% of all concepts in the ontology. This gives urgency to the question of how to find such relevant subsets of concepts from potentially very large vocabularies.

In this paper, we have investigated two methods for identifying such relevant concepts. Our first method consisted of manually identifying a number of seed queries, and performing a lexical search for all concepts whose lexical labels contain any of the seed queries as a substring. All of the resulting concepts and their subconcepts are then considered as relevant for the subdomain characterised by the seed queries followed by the expansion phases.

Our second method consisted of extracting concepts that appeared as a parameter or procedure in the recommendations of the Dutch national guideline for the treatment of breast cancer. These concepts were mapped to SNOMED CT concepts. We compared the results of manual mapping with those obtained using MetaMap tool, which clearly showed that using automatic tool, which focus on lexical matching is not sufficient. This step was followed again by the two expansion phases.

These methods differ from other approaches for the identification of relevant subvocabularies that are based on any *a priori* modularization of the ontology, but instead select sets of concepts that are specific for a particular use of a vocabulary.

*Conclusions.* Our findings indicate that:

- the breast cancer-centric subdomain is indeed only a fraction ($< 1\%$) of all concepts in SNOMED CT
- the seed query method has a high precision, returning only key concepts
- the seed query method has a low recall for returning relevant concepts
- the guideline-method has a higher recall for relevant concepts while still having a high precision for relevant (but possibly non-key) concepts.
- contrary to our prediction, not all key concepts are found by the guideline-method. Close inspection yielded a number of reasons why this is the case in our experiment:
  - the guideline covers only procedures for treatment, hence misses diagnostic concepts
  - we extracted our concepts only from the recommendations in the guideline, hence missing those concepts that only appear in the background information
  - the guideline does not mention procedures that vary between hospitals

*Future Work.* In future work, the validity of our conclusions should be tested by running these experiments on other subdomains (e.g. different diseases), and possibly using other methods to obtain a "gold standard" (our gold standard was obtained by manual extraction of all concepts from a national treatment guideline).

Similarly, it would be interesting to apply the guideline-method to other documents such as patient-records to see if that would yield a very different set of concepts.

More insight should be obtained in the correct choice for the seed concepts, since obviously the method is sensitive to this. The apparent fixed-point behaviour of this method deserves further investigation, for example on the degree of sensitivity to the initial set of query-concepts.

In addition we would like to take a closer look to various modularization algorithms. It would be very interesting to compare modules produced by different methods to learn more about their applicability for identifying disease specific concepts.

## References


1. Aleksovski, Z., Vdovjak, R.: Overlap of selected ontologies in the context of the breast cancer domain. In: Proceedings of SIIM 2009 (2009)
2. Aronson, A.R.: Metamap: Mapping text to the umls metathesaurus. In: Proceedings AMIA Symposium (2001)
3. CBO. Guideline for the Treatment of Breast Carcinoma. van Zuiden. PMID: 12474555 (2002)
4. Clark, K., Parsia, B.: Modularity and owl (2008)
5. Grau, B.C., Horrocks, I., Kazakov, y., Satler, U.: Modular reuse of ontologies: Theory and practise. Journal of Artificial Intelligence Research (2008)
6. Cuenca Grau, B., Horrocks, I., Kazakov, Y., Sattler, U.: Just the right amount: extracting modules from ontologies. In: Proceedings of WWW, pp. 717–726 (2007)
7. Konev, B., Lutz, C., Walther, D., Wolter, F.: Cex and mex: Logical diff and semantic module extraction in a fragment of owl. In: Proceedings of the OWL: Experiences and Directions Workshop, OWLED 2008 (2008)
8. Marcos, M., Galan, J.C., Martinez, B., Polo, C., Seyfang, A., Miksch, S., Serban, R., ten Teije, A., van Harmelen, F., Rosenbrand, K., Wittenberg, J., van Croonenborg, J., Lucas, P., Hommersom, A.: Protocure ii deliverable d2.2bcd: Models of selected guideline in intermediate, asbru and kiv representations. Technical report (2005), `www.protocure.org`
9. McCray, A.T., Srinivasan, S., Browne, A.C.: Lexical methods for managing variation in biomedical terminologies. In: Proceedings of Symposium on Computer Applications in Medical Care, pp. 235–239 (1994)
10. Noy, N.F., Musen, M.A.: The prompt suite: interactive tools for ontology merging and mapping. Int. J. Hum.-Comput. Stud. 59(6), 983–1024 (2003)
11. Porter, M.F.: An algorithm for suffix stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco (1997)
12. Serban, R., ten Teije, A.: Exploiting thesauri knowledge in medical guideline formalization. Methods of Information in Medicine (to appear, 2009)
13. Stuckenschmidt, H., Klein, M.: Structure-based partitioning of large concept hierarchies. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 289–303. Springer, Heidelberg (2004)
14. Suntisrivaraporn, B.: Module extraction and incremental classification: A pragmatic approach for el+ ontologies (2008)